Robust GAN Based on Attention Mechanism

Qian Wu[1,2,3(✉)], Chunjie Cao[1,2,3(✉)], Jianbin Mai[1,2,3], and Fangjian Tao[1,2,3]

[1] Key Laboratory of Internet Information Retrieval of Hainan Province, Haikou 570228, Hainan, China
[2] College of Cryptography, Hainan University, Haikou 570228, Hainan, China
[3] College of Computer and Cyberspace Security, Hainan University, Haikou 570228, Hainan, China

Abstract. Deep neural networks (DNNs) have been found to be easily mislead by adversarial examples that add small perturbations to inputs to produce false results. Different attack and defense strategies have been proposed to better study the security of deep neural networks. But these works only focus on an aspect such as attack or defense. In this work, we propose a robust GAN based on the attention mechanism, which uses the deep latent features of the original image as prior knowledge to generate adversarial examples, and it can jointly optimize the generator and discriminator in the case of adversarial attacks. The generator generates fake images based on the attention mechanism to deceive discriminator, the adversarial attacker perturbs the real images to deceive discriminator, and the discriminator wants to minimize the loss between fake images and adversarial images. Through this training, we can not only improve the quality of adversarial images generated by GAN, but also enhance the robustness of the discriminator under strong adversarial attacks. Experimental results show that our classifier is more robust than Rob-GAN [14], and the generator outperforms Rob-GAN on CIFAR-10.

Keywords: Robust · GAN · Adversarial · Attention mechanism

1 Introduction

In recent years, deep neural networks have achieved great success in image recognition [1], text processing [2], speech recognition [3] and other fields, even widely used in critical security applications, such as malware detection [4], driverless technology [5], aircraft collision avoidance detection [6], etc. These all rely on the security of deep neural networks, which has become the focus of artificial intelligence security. At present, studies have shown that the deep neural network is vulnerable to the disturbance of the original samples with small perturbations [7]. These disturbances can make the system produce wrong judgment results while cannot be perceived by human eyes. Such input samples are called adversarial samples [8]. Adversarial examples can not only pose potential threat by attacking deep neural networks, but also enhance the robustness of models through training models [9]. Therefore, it is necessary to study the generation of adversarial samples.

J. Cheng et al. (Eds.): CSS 2020, LNCS 12653, pp. 78–86, 2021.
https://doi.org/10.1007/978-3-030-73671-2_8

Adversarial samples can be divided into two categories according to the attack target: maliciously-chosen target class (targeted attack) or classes that are different from the ground truth (non-targeted attack). At present, different methods have been proposed to generate adversarial samples. These methods are mainly divided into three categories. The first is gradient-based attack, such as the Fast Gradient Sign Method (FGSM) [8], which uses the linear nature of the deep neural network model in the high-dimensional space to quickly obtain the anti-perturbation, and adds disturbances in the gradient direction of the input vector. However, there is a minimization problem in this way. The second is optimization-based attack. Such as C&W attack [10], by limiting the distance l_0, l_2, l_∞ norms from the real image, the perturbation amplitude of the adversarial sample is reduced. But this method is slow because it can only focus on one instance at a time. The third is generative-network based attack. Such as Natural GAN [11], which generates adversarial examples of text and images by GAN and makes the generated adversarial examples more natural. These methods are also used in black box attack. Although the generation speed of these methods is fast, the disturbance is usually larger than the above two types of methods, and it's easy to be found.

Contrary to adversarial attacks, adversarial defenses are techniques that enable the model to resist adversarial samples. Compared with attacks, defenses are more difficult. Nevertheless, a large number of defense methods are still proposed, mainly in two aspects: the passive defenses, including input reconstruction, confrontation detection, and the active defenses, including defense distillation [12] and adversarial training [13].

However, the researches in these networks only focus on one aspect of attack or defense, and do not consider improving attack and defense simultaneously within a framework.

Our contribution in this work is:

A robust generative adversarial network based on the attention mechanism (Atten-Rob-GAN) is proposed. By introducing the attention mechanism to extract the original image features and use them as the input of generator G, the network can learn the relationship between the deep features of the image. Fake images generated by G are inputted into the discriminator D, while the adversarial images obtained from the attacker interference with the original images are also inputted into D. The adversarial training and GAN training are coordinated to obtain a powerful classifier, while improving the training speed of GAN and the quality of the generated images.

2 Materials and Methods

In this section, we will first introduce the definition of the problem, then briefly describe the framework of the Atten-Rob-GAN algorithm, and the method used to generate attacked images, finally explain the network in detail, concluding the formula and training details used in our framework.

2.1 Problem Definition

$x \in R^n$ is the original sample feature space, and n is the feature dimension. (x_i, y_i) is the i-th instance in the training set, which is composed of a feature vector $x_i \in X$

generated from an unknown distribution $x_i \sim P_{real}$ and the corresponding ground truth label $y_i \in Y$. Let $x_{fake} \in R^n$ be the feature space of false sample, and n is the feature dimension. (x_{fake_i}, l_i) is the i-th sample pair in the false sample data set, x_{fake_i} obeys an unknown distribution P_{fake}, and l_i is the corresponding prediction label. x_{adv} is the original image preprocessed by the PGD attack. The discriminator encourages x_{fake} to approximate x_{adv} within the perturbation range, so that P_{fake} is close to P_{real}.

2.2 The Atten-Rob-GAN Framework

Figure 1 shows the overall framework of Atten-Rob-GAN, which mainly includes three parts: feature extractor F, generator network G, and discriminator D. The output $F(x)$ of the feature extractor F which input is the real image and the noise vector z are concatenated vectors to form $F(x)$*. The generator G receives $F(x)$* to generates the fake image x_{fake}. The discriminator D receives the image x_{adv} and the generator output x_{fake}, and distinguishes them, predict the category when the judgment is true.

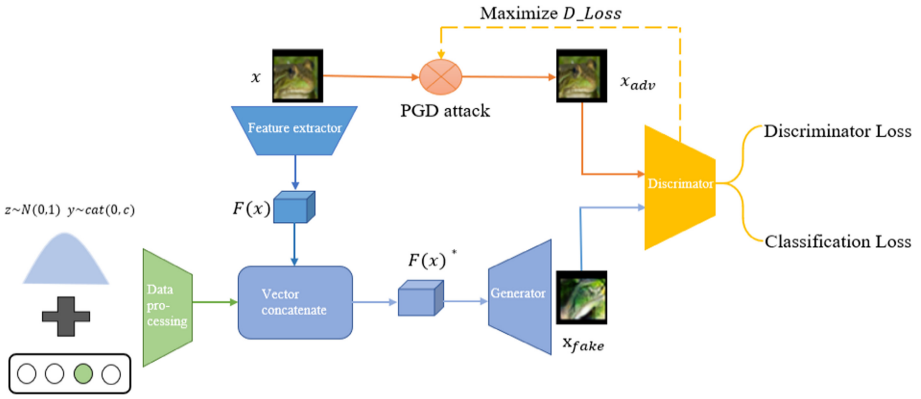

Fig. 1. The network architecture

The Loss Function

This work uses the same loss function as in Rob-GAN [14], the discriminator judges the source and category of the image, $P(S|X)$, $P(C|X) = D(X)$. The only difference is that the generator G adds the deep features of the original image for feature fusion as input, $X_{fake} = G((c, z) + F(X_{real}))$. The loss function has two parts:

Discriminator Loss:

$$L_s = E\big[\log P(S = real|X_{real})\big] + E[logP(S = fake|X_{fake})] \tag{1}$$

Classification Loss:

$$L_{c_{real}} = E[logP(C = c|X_{real})] \tag{2}$$

$$L_{c_{fake}} = E[logP(C = c|X_{fake})] \tag{3}$$

Train the discriminator D to maximize $L_s + L_{c_{real}}$, and train the generator G to minimize $L_s - L_{c_{fake}}$.

2.3 The Method of Generating Adversarial Examples Datasets

Projected Gradient Descent (PGD)
Madry et al. proposed an attack used in adversarial training called "Projected Gradient Descent" (PGD) [15] in 2017. Here, the PGD attack refers to initializing a search for an adversarial instance at a random point within the allowed norm sphere, and then running several basic iterative methods [16] to find adversarial examples. Given an example x, whose ground truth label is y, the PGD attack calculates the adversarial disturbance δ by using the projection gradient descent to solve the following optimization:

$$\delta := \underset{||\delta|| \leq \delta_{max}}{argmax} l(f(x+\delta; w), y) \tag{4}$$

Where $f(.; w)$ is the network parameterized by the weight w, $l(., .)$ is the loss function, and we choose $||.||$ as the l_∞ norm. The PGD attack is the strongest attack in first-order gradient attack. Using this attack to conduct adversarial training will make the defense more successful.

2.4 Implementation

Network Architecture
Next, we briefly introduce the network structure of Atten-Rob-GAN. For a fair comparison, we copied all the network architectures of the generator and discriminator from Rob-GAN. Other important factors, such as learning rate, optimization algorithm, and the number of discriminator updates in each cycle also remain unchanged. The only modification is that we added an attention mechanism to the input of the generator, the feature extractor (see Fig. 3).

Generator
The specific network structure of the generator is shown in Table 1:

Table 1. The specific structure of Atten-Rob-GAN generator

Layers	Types	Input channel number	Output channel number	Activation function	Up-sample
1	Linear	128	64×16		
2	Block1	64×16	64×8	ReLU	True
3	Block2	64×8	64×4	ReLU	True
4	Block3	64×4	64×2	ReLU	True
5	Block4	64×2	64×1	ReLU	False
6	BatchNorm2d	64	64	ReLU	
7	Conv2d (3 * 3)	64	3	tanh	

The first layer of the generator is a fully connected layer that the input is 128 noise, and the output is a $4^2 \times 64 \times 16$ image, where 4^2 is the size of the feature map, and 64

× 16 is the number of channels. Then there are 4 residual blocks, a batch regularization, and the last layer is a convolutional layer with the size of a 3 * 3 convolution kernel.

Discriminator
The specific network structure of the discriminator is shown in Table 2:

Table 2. The specific structure of Atten-Rob-GAN discriminator

Layers	Types	Input channel number	Output channel number	Activation function	Down-sample
1	Optimized block	3	64 × 2		True
2	Block2	64 × 2	64 × 2	ReLU	True
3	Block3	64 × 2	64 × 2	ReLU	False
4	Block4	64 × 2	64 × 2	ReLU	False
5	Activation	64 × 2	64 × 2	ReLU	
6	Linear	64 × 2	1 (sources) 10 (classes)		

The first layer of the discriminator is the optimized residual block, its detailed information is shown in Fig. 2. Then there are 3 residual blocks, an activation layer, a fully connected layer. The last fully connected layer has two types, in one case, the number of output channels is 1 when judging true or false image, and the other is that the number of output channels is the number of categories when judging the image category.

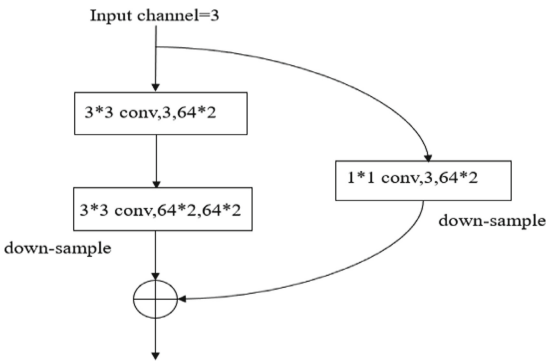

Fig. 2. Optimized block

Feature Extractor Based on Attention Mechanism

Here, we first extract the image features by reducing the dimension of the original image through a network structure completely symmetrical to the generator network, then introduce the attention mechanism (SE module [17]) to extract the spatial relationship in the image's shallow features and channel feature relationship to form deep features, so that the image can learn the weight coefficients of different channel features, thus the model can make more discerning about the characteristics of each channel. Figure 3 shows the detailed process of feature extractor F.

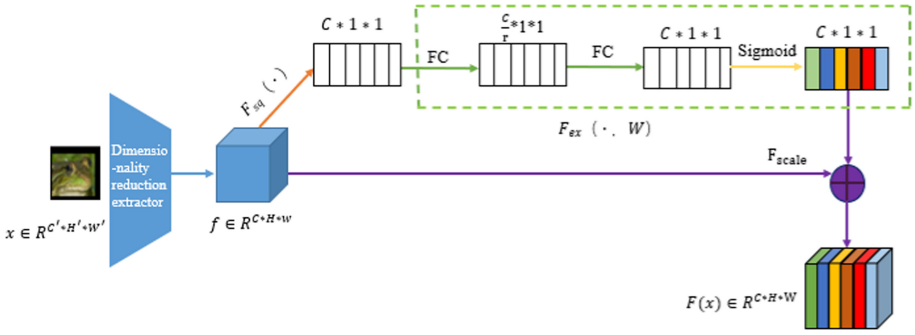

Fig. 3. Feature extraction (SE [17])

Training Details

We conduct experiments on the MNIST [18] and CIFAR-10 [19], where we use the training set to train Rob-GAN and Atten-Rob-GAN respectively, and evaluate the test set. After the model training is completed, the test set is input to the discriminator for testing, and the accuracy of the model is used as the measurement standard. The Adam optimizer with a learning rate of 0.0002 and $\beta_1 = 0, \beta_2 = 0.9$ is used to optimize the generator and discriminator. We sample the noise vector from the normal distribution and use label smoothing to stabilize the training.

Implementation Details

In our experiment, we use Pytorch for implementation and run on NVIDIA GeForce RTX 2080 Ti * 2. We train Atten-Rob-GAN to be 200 eopchs, batch size is 64, learning rate is 0.0002, attenuation by 50% every 50 steps, and PGD attack intensity is assumed to be 0.0625.

3 Results and Discussion

3.1 Robustness of Discriminator

In this experiment, we compared the robustness of the discriminator trained by Atten-Rob-GAN with Rob-GAN. As shown in [14], the robustness of Rob-GAN under adversarial attacks even surpasses the state-of-the-art adversarial training algorithm [15]. In

the comparison of [20], adversarial training was considered to be the latest level of robustness. Since Rob-GAN is equivalent to Atten-Rob-GAN without an attention mechanism component to extract feature, for fair comparison, we keep all other components the same. In order to test the robustness of the model, we choose the widely used l_∞ PGD attack [15], but using other gradient-based attacks is also expected to produce the same results. As defined in (8), we set l_∞ disturbance as $\delta_{max} \in np.range(0, 0.01, 0.02, 0.03, 0.04)$. In addition, we scale the image to $[-1, 1]$ instead of $[0, 1]$, because the last layer of the generator has $tanh()$ output, so we need to modify it accordingly. We display the results in Table 3, all results are the average results after 5 runs.

Table 3. Accuracy of our model under l_∞ PGD-attack.

Datasets	Defenses	δ_{max} *of* l_∞ attack				
		0	0.01	0.02	0.03	0.04
CIFAR 10	Rob-GAN [14]	78.99%	69.54%	58.47%	50.78%	35.51%
	Ours	**88.37%**	**79.49%**	**66.94%**	**51.53%**	**36.35%**
MNIST	Rob-GAN [14]	52.43%	50.59%	49.37%	47.35%	45.05%
	Ours	**61.31%**	**57.32%**	**53.92%**	**49.20%**	44.91%

We can observe from Table 3 that our model has a higher classification success rate than Rob-GAN without attack, which proves that our classifier is more accurate after training. At the same time, under the attack intensity of [0, 0.04], our accuracy is higher than Rob-GAN's classifier on CIFAR-10, which proves that our model can obtain a more robust classifier. In the case of an attack intensity of 0.04 on MNIST, our result is slightly lower than that of Rob-GAN. The reason may be that the number of experiments is too few, and the calculated mean result is not universal.

3.2 Quality of Generator

Finally, we evaluate the quality of the generator trained on the CIFAR-10 dataset by comparing it with the generator obtained by Rob-GAN. Figure 4 shows the adversarial images generated on the two models. We can clearly observe that the image quality generated by Atten-Rob-GAN is significantly better than Rob-GAN, and even brighter than the original image.

(a) The real images (b) Images generated by our model (c) Images generated by Rob-GAN

Fig. 4. Different generated images.

4 Conclusion

We propose a robust generative adversarial network based on the attention mechanism. By adding the attention mechanism, the features of the original image can be extracted deeply, thereby improving the quality of the image generated by the generator. At the same time, the discriminator and generator are jointly trained in the case of adversarial attack to obtain a more powerful discriminator, this method can effectively improve the robustness of the classifier. And through experimental comparison, it is proved that the attention mechanism component we added has an optimization effect on Rob-GAN, both in terms of the robustness of the discriminator and the quality of the generator.

Acknowledgements. We acknowledge the support by the National Natural Science Foundation of China (No. 66162019); National Natural Science Foundation of China Enterprise Innovation and Development Joint Fund (No. U19B2044).

References

1. Krizhevsky, I.S., Hinton, G.E.: ImageNet classification with deep convolutional neural networks. Commun. ACM **60**(6), 84–90 (2017)
2. Bahdanau, D., Cho, K., Bengio, Y.: Neural machine translation by jointly learning to align and translate. In: 3rd International Conference on Learning Representations, pp. 11–17. ICLR, San Diego (2015)
3. Hinton, G., Deng, L., Yu, D., et al.: Deep neural networks for acoustic modeling in speech recognition: the shared views of four research groups. IEEE Sig. Process. Mag. **29**(6), 82–97 (2012)
4. Yuan, Z., Lu, Y.Q., Wang, Z.G., Xue, Y.B.: Droid-Sec: deep learning in android malware detection. In: ACM SIGCOMM 2014 Conference, pp. 371–372. ACM, Chicago (2014)
5. Eykholt, K., Evtimov, I., Fernandes, E., et al.: Robust physical world attacks on deep learning models. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1625–1634. IEEE Computer Society, Salt Lake City (2018)
6. Majumdar, R., Kunčak, V. (eds.): CAV 2017. LNCS, vol. 10426. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-63387-9
7. Cubuk, E.D., Zoph, B., Schoenholz, S.S., Le, Q.V.: Intriguing properties of adversarial examples. In: 6th International Conference on Learning Representations (ICLR), pp. 106–118. ICLR, Vancouver (2018)

8. Goodfellow, I.J., Shlens, J., Szegedy, C.: Explaining and harnessing adversarial examples. In: 3rd International Conference on Learning Representations, pp. 65–78. ICLR, San Diego (2015)
9. Tramèr, F., Kurakin, A., Papernot, N., GoodfellowI, B.D., McDaniel, P.: Ensemble adversarial training: attacks and defences. In: 5th International Conference on Learning Representations, pp. 123–142. ICLR, Toulon (2017)
10. Carlini, N., Wagner, D.: Towards evaluating the robustness of neural networks. In: IEEE Symposium on Security and Privacy, vol. 0, pp. 39–57. IEEE, San Jose (2017)
11. Zhao, Z., Dua, D., Singh, S.: Generating natural adversarial examples. In: 6th International Conference on Learning Representations, pp. 108–115. ICLR, Vancouver (2018)
12. Papernot, N., Mc Daniel, P., Wu, X., Jha, S., Swami, A.: Distillation as a defense to adversarial perturbations against deep neural networks. In: 2016 IEEE Symposium on Security and Privacy (SP), pp. 582–597. IEEE, San Jose (2016)
13. Wu, Y., Bamman, D., Russell, S.: Adversarial training for relation extraction. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pp. 1778–1783. ACL (2017)
14. Liu, X., Hsieh, C.: Rob-GAN: generator, discriminator, and adversarial attacker. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 11226–11235. IEEE Computer Society, Long Beach (2019)
15. Madry, A., Makelov, A., Schmidt, L., Tsipras, D., Vladu, A.: Towards deep learning models resistant to adversarial attacks. In: 5th International Conference on Learning Representations, pp. 1538–1549. ICLR, Toulon (2017)
16. Kurakin, A., Goodfellow, I., Bengio, S.: Adversarial examples in the physical world. In: 5th International Conference on Learning Representations, pp. 995–1012. ICLR, Toulon (2017)
17. Hu, J., Li, S., Albanie, S., Sun, G., Wu, E.: Squeeze-and-excitation networks. IEEE Trans. Pattern Anal. Mach. Intell. **42**(8), 2011–2023 (2018)
18. LeCun, Y., Cortes, C.: MNIST handwritten digit database. Proc. IEEE **86**(11), 2278–2324 (1989)
19. Krizhevsky, A., Nair, V., Hinton, G.: CIFAR-10 (Canadian institute for advanced research). http://www.cs.toronto.edu/~kriz/cifar.html
20. Athalye, A., Carlini, N., Wagner, D.: Obfuscated gradients give a false sense of security: circumventing defenses to adversarial examples. In: 35th International Conference on Machine Learning, ICML 2018, vol. 1, pp. 436–448. IMLS, Stockholm (2018)